Journal of Biomedical Informatics 50 (2014) 151–161

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

De-identification of clinical notes in French: towards a protocol for reference corpus development

Cyril Grouin *, Aurélie Névéol

LIMSI–CNRS, UPR 3251, Orsay, France

ARTICLE INFO

Article history:
Received 2 August 2013
Accepted 22 December 2013
Available online 29 December 2013

Keywords:
Confidentiality
Electronic Health Records
France
Information Dissemination
Natural Language Processing

ABSTRACT

Background: To facilitate research applying Natural Language Processing to clinical documents, tools and resources are needed for the automatic de-identification of Electronic Health Records.
Objective: This study investigates methods for developing a high-quality reference corpus for the de-identification of clinical documents in French.
Methods: A corpus comprising a variety of clinical document types covering several medical specialties was pre-processed with two automatic de-identification systems from the MEDINA suite of tools: a rule-based system and a system using Conditional Random Fields (CRF). The pre-annotated documents were revised by two human annotators trained to mark ten categories of Protected Health Information (PHI). The human annotators worked independently and were blind to the system that produced the pre-annotations they were revising. The best pre-annotation system was applied to another random selection of 100 documents. After revision by one annotator, this set was used to train a statistical de-identification system.
Results: Two gold standard sets of 100 documents were created based on the consensus of two human revisions of the automatic pre-annotations. The annotation experiment showed that (i) automatic pre-annotation obtained with the rule-based system performed better ($F = 0.813$) than the CRF system ($F = 0.519$), (ii) the human annotators spent more time revising the pre-annotations obtained with the rule-based system (from 102 to 160 minutes for 50 documents), compared to the CRF system (from 93 to 142 minutes for 50 documents), (iii) the quality of human annotation is higher when pre-annotations are obtained with the rule-based system (*F*-measure ranging from 0.970 to 0.987), compared to the CRF system (*F*-measure ranging from 0.914 to 0.981). Finally, only 20 documents from the training set were needed for the statistical system to outperform the pre-annotation systems that were trained on corpora from a medical speciality and hospital different from those in the reference corpus developed herein.
Conclusion: We find that better pre-annotations increase the quality of the reference corpus but require more revision time. A statistical de-identification method outperforms our rule-based system when as little as 20 custom training documents are available.

1. Introduction

Medical knowledge is routinely advanced through clinical studies involving patient volunteers who provide informed consent to participate in a carefully designed study, planned before any medical information is collected or any health care procedure is performed. Medical knowledge can also be greatly advanced through retrospective studies exploiting the wealth of information contained in Electronic Health Records (EHRs). This type of study also requires the patients involved to provide informed consent. However, because the study design is crafted after the patients have received the care described in the EHRs, it can be difficult to obtain consent from each patient (e.g., logistics issues arise for contacting the patients or their family).

De-identification is the process of hiding or removing content that explicitly identifies persons involved in patient care, including patients themselves and health care providers [1]. The use of de-identified clinical data provides researchers with the means to carry out studies that can advance the state of medical knowledge while protecting patients' privacy and confidentiality. Specifically, in the absence of informed consent, the Personally Identifiable Information (PII) and Protected Health Information (PHI) contained in clinical data must be processed according to privacy rules and regulations.

A significant body of research has addressed the issue of de-identification in the past decades, covering different types of data,

* Corresponding author.
E-mail addresses: cyril.grouin@limsi.fr (C. Grouin), aurelie.neveol@limsi.fr (A. Névéol).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.12.014

such as text, images, biological samples and DNA sequences [2]. In this paper, we focus on the de-identification of clinical free-text, as a preliminary step to prepare clinical text for further Natural Language Processing (NLP) and analysis of clinical documents. In order to ensure the quality and robustness of NLP tools, real clinical data must be used for development and testing.

An increasing number of efforts recently targeted the de-identification of clinical text in English [3]. Other efforts also addressed the de-identification of clinical documents in languages other than English such as French [4,5] and Swedish [6,7]. The lack of a freely available de-identification reference corpus similar to the i2b2 corpus available for English [8] has prevented any rigorous comparison between the two approaches developed for French.

Our goal is to support research in Natural Language Processing for biomedical texts in French through the development of a de-identified corpus that can be distributed to the scientific community for research purposes [9].

In this paper, we focus on three specific aims that address both fundamental research questions and practical considerations:

1. De-identification research methods for clinical texts in French: what are the best methods for automatic text de-identification in French? What are the best methods for producing a reliable, high-quality reference corpus for de-identification? Specifically, we assess the usability of two automatic pre-annotation methods.
2. De-identification resources: development of a de-identification reference corpus freely available to the scientific community.
3. De-identification evaluation: assess the time and effort required to produce de-identified corpora and adapt existing de-identification tools to new, unseen data.

2. Related work

2.1. De-identification of clinical free-text

De-identification of clinical data, including de-identification of clinical free-text in English has been well-studied in the past decade. De-identification is generally approached as a specific named entity recognition task targetting PHIs. Named entity recognition is defined by Meystre et al. [1] as "the task of recognizing expressions denoting entities (i.e., named entities), such as diseases, drugs, or people's names, in free text documents". A review of available tools shows that de-identification can be reasonably achieved using a rule-based approach, statistical machine learning, or a combination of both [3]. The rule-based tool developed by Neamatullah et al. [10] was notably used to de-identify clinical documents in the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database [11,12] and adapted to data outside of the United States [13]. It uses the principle of surrogate PHI re-introduction, which consists of substituting PHI in the original records by similar made-up data in order to preserve language coherence while enforcing privacy. This process was shown to have minimal impact on information extraction in clinical documents [14].

Recent work used the de-identification reference corpus developed for the i2b2 2006 challenge [8] to perform a systematic evaluation of five de-identification systems available for English [15], which prompted the development of a new tool customized for VA documents [16]. Contrary to what was reported by Wellner et al. [17] in their work for the i2b2 challenge, these studies showed that a fair amount of adaptation is required for any de-identification tool to obtain acceptable results on new, unseen corpus. The study also provided valuable insight for de-identification tool adaptation by pointing out that the strength and weaknesses of rule-based and statistical systems seem to be different for the types of PHI targeted.

While de-identification as a task seems to be almost resolved, efficient adaptation of de-identification tools to new corpora (including in languages other than English) currently remains a major challenge. Recent work provided estimations of the annotator time [18,19] required to prepare training data for a statistical de-identification tool achieving 0.96 *F*-measure on clinical notes in English. The same group also assessed the effect of training corpus size [19], training document type [18] and re-identification status [20].

2.2. Development of annotated reference corpora

Many international NLP challenges require annotated reference corpora for participant evaluation. For instance, the Message Understanding Conferences (MUC) [21] notably produced reference corpora for named entities such as organization names, person names, locations, dates and times in newswire text. In the biomedical domain, challenges yielded reference corpus for named entities including bacteria [22], genes [23], medications [24], diseases [25] or PHI [8]. Throughout the tasks, a variety of document genres were covered, including scientific articles or abstracts [22,23,25], clinical text [8,24], PubMed queries [26]. Several methods have been used and assessed for producing reference annotations for these tasks, relying on human annotations for all [8,24,25] or part [23] of the final reference. Automatic pre-annotations have often been used to process corpora in the biomedical domain in order to create quality reference corpora [8,26,27]. This process was shown to be relevant and useful as it saves annotation time, contributes to the production of consistent, high quality annotations and is overall preferred by annotators [26]. Another commonly used method for producing high-quality reference corpora is the use of several annotators working either independently [25,27] or in sequence [8] and discussing disagreements to produce consensus annotations.

3. Material and methods

3.1. Corpus

The corpus we used was approved by the French administrative authority on data privacy[1] for research on Information Retrieval (IR) in large Electronic Health Records [28]. To address the IR task in the context of severe diseases (i.e., records containing a large number of documents on a given patient) 1000 patient records were selected randomly from patients with at least 50 hospital stays in a group of hospitals within a French geographic area. The entire corpus comprises about 170,000 documents. As a result, a large variety of medical specialties (e.g., Pneumology, Obstetrics, Infectious Diseases), clinical document types (e.g., radiography reports, discharge summaries, consult correspondance) and hospitals (5 locations) are covered in the corpus. In the random subsets of documents used in this study, we did not attempt to control the distribution of either specialties, document types, or original health care provider. The sheer number of document types and specialties represented in the overall corpus would make this a difficult task. While it has been shown that tools perform better if trained on documents very similar to those they are tested on [18], we are interested in assessing the portability of de-identification tools with minimal adaptation work.

The corpus was de-identified by the original health care providers based on patient information as it appeared in the local hospital information system: patients' first and last names were replaced by the string "XX" while the day and month in their dates of birth

[1] Commission Nationale de l'Informatique et des Libertés (CNIL).

Original	Nom : XX Prénom : XX né le : jj mm.1932 N° Dossier ..... Date d'entrée : 31/07/2013 Date de sortie : 07/10/2013 No Loghos : 012345678
Translation	*Last name: XX First name: XX* *Born: dd mm.1932 File Nb .....* *Entry Date: 2013/07/31 Discharge Date: 2013/10/07* *Loghos Id: 012345678*

Fig. 1. Excerpt from the partially de-identified corpus. Last name, first name and date of birth of the patient have been replaced by a generic string; file number has been deleted.

were replaced by the string "jj mm".[2] The birth year remained verbatim (see Fig. 1). In addition, a baseline method was used to scrub patient record numbers from the documents. When the phrase "N° dossier" (record number) appeared followed by an identification number, the identification number was removed.

In the remainder of the paper, all excerpts from the corpus used for illustration purposes contain surrogate PHI.

3.2. Annotation guidelines

We defined guidelines for marking PHI in clinical documents.

The guidelines list the categories to annotate and provide examples for each category. Ten categories[3] were defined based on the United States Health Insurance Portability and Accountability Act (HIPAA) "Safe Harbor" [29], the type of information likely to be found in the corpus and previous annotation experiments conducted on clinical documents in French [5]:

- Person names:
  - *lastname*: last names of patients, family members or healthcare providers—except in street names or health care provider locations where the "address" or "hospital" categories must be used;
  - *firstname*: first names of patients, family members or healthcare providers.
- Dates:
  - *date*: all dates, including information on day, month and year when available (*01.01.2012; 2 mars 2013; Sept. 2009; 04/11*).
- Places:
  - *hospital*: health care provider locations such as names of hospitals or specific wards within a hospital (*Hôpital de Saint-Nazaire, Unité J. Dupont*);
  - *address*: postal address, including street address or specific buildings within a hospital (*1, rue de Paris; cour Martin*);
  - *zipcode*: zip code (*75013*);
  - *city*: city name, except in street names or health care provider locations where the "address" or "hospital" categories must be used (*Saint-Nazaire; Paris*).
- Contact data and identifiers:
  - *telephone*: telephone number (*01.23.45.67.89*), fax number and telephone extension without the trigger word (*poste 12345*);
  - *email*: electronic mail address (nom.prenom@hopital-ville.fr);
  - *id*: all identifiers, either numerical or alphabetical (patient record number, social security number, hospital ward identification number, device serial number, etc.).

3.3. Annotation methodology

Fig. 2 presents the overall annotation methodology. We worked sequentially with three sets of 100 documents. We adjusted the annotation methodology from one set to the next in order to optimize the annotation task in terms of annotation quality, annotation time, annotator experience.

Two annotators participated in the study (the authors of this paper). The annotators had extensive previous experience with annotations in the biomedical domain (AN, CG) as well as the general domain (CG).

We randomly extracted 100 documents from the corpus (Set 1). A first subset of 10 documents was used to test the principles of the annotation task outlined in the guidelines and to allow the human annotators to become familiar with the annotation tool. These documents were used for training the annotators to perform the task without any influence from the pre-annotations. In this stage, annotators were encouraged to discuss any aspect of the annotation task. In subsequent stages of the annotation process, the annotators worked independently. A second subset of 10 documents was set aside to be annotated from scratch, i.e., no pre-annotations were provided to the annotators. This small set of 10 documents was used in the revision phase described below in order to allow for a comparison of annotations obtained on the same documents with and without access to pre-annotations.

Then, the entire set of 100 documents was automatically pre-annotated as outlined below, including the 10 documents that were already annotated from scratch. The annotators revised the pre-annotated corpus. Finally, both annotators reviewed the cases where they provided different annotations and agreed to a final consensus annotation. Previous studies revealed that human annotators tend to trust automatic pre-annotation and make little changes to system outputs [26,30]. Therefore, a reviewed consensus set of annotations can be used a high-quality gold standard against which to compare both human and automatic annotations.

Based on the evaluation of the pre-annotations on the first set of 100 documents using the consensus annotations, a second set of 100 randomly extracted documents (Set 2) was pre-annotated with the rule-based system and revised by one annotator (AN). This set was used in further experiments to train a statistical system adapted to the consensus set, used as a test set.

Based on the evaluation of the custom-trained statistical system, it was used to pre-annotated a third set of 100 randomly selected documents (Set 3). This set was revised by both annotators and consensus annotations were produced.

3.3.1. Pre-processing

In order to process real data, we reintroduced realistic surrogate information in the corpus [31]. For instance, the phrase "jj mm" appearing within patients' dates of birth was replaced with "*01.01*" or "*01/01*" (a default *January 1st* value). The symbol used to separate the year from "jj mm" was kept to separate the reintroduced day and month (e.g., "*jj mm/2013*" became "*01/01/2013*" while "*jj mm.2013*" became "*01.01.2013*").

The "XX" phrases appear in the corpus in several contexts: (i) a trigger word followed by two occurrences of the phrase: *Mme XX XX* (Ms. XX XX), (ii) a trigger word followed by one occurrence of the phrase: *M. XX* (Mr. XX), and (iii) a single occurrence of the phrase without any trigger word: *XX*.

[2] "jj mm" stands for "*jours mois*" in French, i.e., "dd mm" in English.

[3] There is no direct equivalence between our ten categories and the 18 HIPAA identifiers. Some HIPAA identifiers were grouped into one category in our annotation schema (e.g., social security numbers and records numbers), other identifiers were split into two categories (e.g., first names and last names). Finally, some identifiers were not relevant for our corpus (e.g., vehicle identifiers and serial numbers) and study (biometric identifiers and full face photographic images).

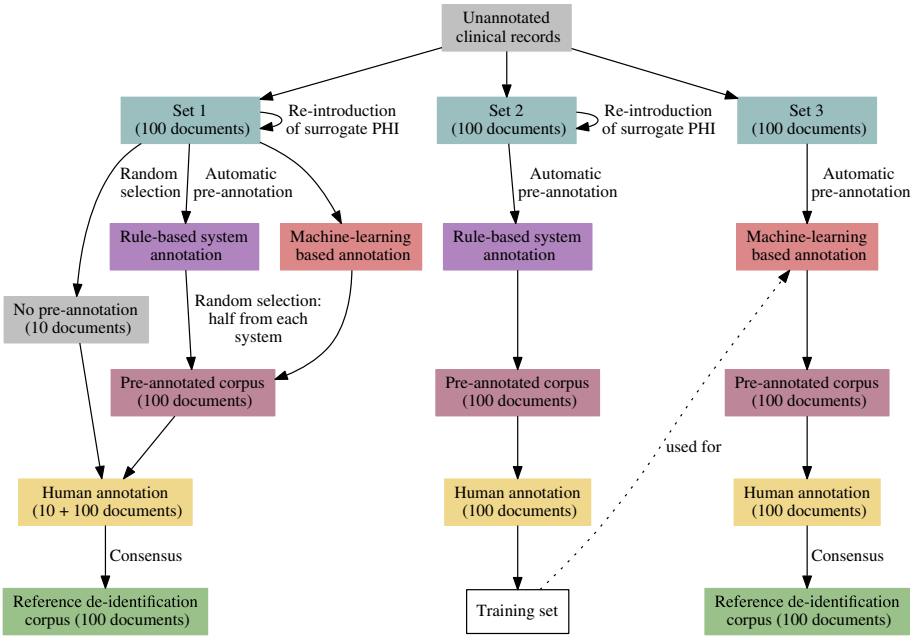

Fig. 2. Overall annotation methodology.

Depending on the context, the "XX" phrases were replaced either by a first name or a last name. We used a last name followed by a first name to replace two consecutive phrases, a last name to replace one phrase when it followed a trigger word, and a last name to replace one phrase when it was used without any trigger word. The first names and last names used were randomly selected from lists of international and French names. The lists of names are freely available from the French ABU association website.[4] The lists were pre-processed to remove first names and last names that are not used in France but are similar to existing common French words (e.g. Agace, Cela, Le, Travers) and are likely to produce false positives.

3.3.2. Automatic pre-annotations

We used the MEDINA suite of tools to produce automatic pre-annotations. It comprises two de-identification systems that have been designed to process clinical notes from a cardiology ward written in French [5,31]. As shown in Fig. 3 the systems output text-bound annotations as tags surrounding the phrases to be de-identified.

The first system is a rule-based system [31] that relies on 80 patterns specifically designed to process the training corpus and lists[5] we gathered from existing resources from the Internet. The system implements several steps: (i) identification of numeric data with patterns,[6] (ii) lexicon mapping (exact match using lexicons of first names, last names, city names, etc.), (iii) identification of named entity with trigger words (e.g., "M." and "Mme" are trigger words for last names in French) and patterns, and (iv) study of the neighborhood of already processed data.

The second system is based on the CRF formalism [32] as implemented in Wapiti [33]. We used four types of features to build our models [5]: (i) surface features (token, capitalization, digit, punctuation, token length), (ii) morpho-syntactic features (token part-of-speech and surrounding tokens POS), (iii) semantic types (based on previous lexicon, and token Concept Unique Identifier from the UMLS Metathesaurus), and (iv) distributional analysis through an unsupervised clustering based on the Brown algorithm [34] as implemented in Liang's tool [35]. We did not perform cross-validation, but automatic feature selection was carried out through the $l1$ regularization.

Both systems were used to process the 100 documents selected in Set 1. For each document in the set, we randomly selected one output to be shown to the annotators for revision. As a result, the corpus provided to the annotators for revision comprised 100 documents where 50 had been processed by the rule-based system and 50 had been processed by the CRF-based system. The annotators were blind to the type of pre-annotation performed on the documents they reviewed.

The rule-based system was used to process Set 2. One annotator reviewed the output. The CRF system was re-trained on Set 2 using the original set of features and used to process Set 3. Both annotators independently reviewed the documents in this set.

Before revision, the corpus presented to the annotators comprised a total of 2900 automatically produced pre-annotations for 100 documents (29,437 tokens) in Set 1. Fig. 4 shows the distribution of annotations over the PHI category types in the corpus. Table 1 shows the variability of the occurrences per category in the reference corpus (e.g., while 99 addresses were marked in total, there were only 29 distinct addresses occurring in the corpus). We observed a similar distribution in Set 2 and Set 3.

3.3.3. Annotation tool

Human annotations were produced using the Brat annotation tool[7] [36]. This tool is commonly used to annotate entities and relations between entities in biomedical corpora. For instance, it was used a supporting tool for international challenges such as BioNLP Shared Task 2011 and 2013, CoNLL 2000, 2002 and 2006, Drug-Drug Interaction 2011. In a recent survey of annotation tools for the biomedical domain, BRAT was reported to be easy to install, user friendly and fast [37]. For the purpose of the present study we developed two perl scripts that convert text with embedded annotation tags (produced by our automatic pre-annotation tools and used by the tool that we used to compute inter-annotator agreement) to

[4] http://abu.cnam.fr/DICO/, Association des Bibliophiles Universels, CNAM, Paris.

[5] The lists we used contain 23,000 first names, 12,800 last names, 30,700 city names, 2000 hospital names, and 251,000 inflected forms from the French language.

[6] e.g., the pattern "1-2 DIGIT (S) SEPARATOR 1-2 DIGIT (S) SEPARATOR 2-4 DIGITS" is a rule that extracts dates such as "01/01/2014" or "1-12-88".

[7] Brat Rapid Annotation Tool, http://brat.nlplab.org/.

COMPTE - RENDU D'HOSPITALISATION :
Nom: <nom>Le Rouge</nom> Prénom: <prenom>Rackham</prenom>
Né (e) le : <date>01/01/1968</date>

Fig. 3. Sample output from a pre-annotation tool.

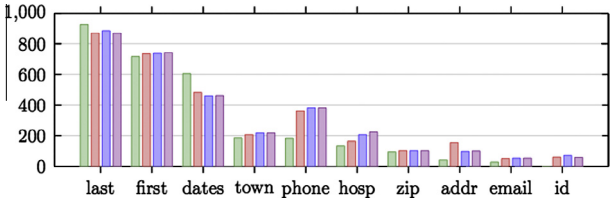

Fig. 4. Number of annotations per category: automatic pre-annotations (green), first annotator (red), second annotator (blue), consensus (purple). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and from the BRAT stand-alone format. The scripts are freely available from the authors upon request.

3.4. Evaluation metrics

3.4.1. Inter- and intra-annotator agreements

We computed inter-annotator agreement on the raw corpus (10 documents from Set 1) and on the pre-annotated corpora (Set 1 and Set 3). Two agreement metrics were used, namely the κ coefficient and the *F*-measure.

We used the κ coefficient defined by Cohen [38] (formula (1), where A_o stands for the observed agreement[8] and A_e stands for the expected agreement,[9] computed using formula (2). In formula (2), i stands for the number of categories and n_k is the number of annotations for category k).

$$\kappa = \frac{A_o - A_e}{1 - A_e} \tag{1}$$

$$A_e^{\kappa} = \sum_k \frac{nA_1k}{i} \times \frac{nA_2k}{i} \tag{2}$$

We also used *F*-measure (formula (7)), which is the weighted harmonic mean between recall[10] (formula (3)) and precision[11] (formula (5)). Specifically, we used $\beta = 1$. Micro-average (formulae (4) and (6) where i stands for the number of categories) was used to compute overall results on all categories [39].

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{3}$$

$$\text{Micro-recall} = \frac{\sum_{i=1}^{n} \text{true positive}(i)}{\sum_{i=1}^{n} \text{true positive}(i) + \sum_{i=1}^{n} \text{false negative}(i)} \tag{4}$$

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{5}$$

$$\text{Micro-precision} = \frac{\sum_{i=1}^{n} \text{true positive}(i)}{\sum_{i=1}^{n} \text{true positive}(i) + \sum_{i=1}^{n} \text{false positive}(i)} \tag{6}$$

$$F\text{-measure} = \frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \tag{7}$$

[8] The observed agreement corresponds to the number of common annotations between both annotators.
[9] The expected agreement is the hypothetical probability of chance agreement.
[10] Recall, or true positive rate, or sensitivity.
[11] Precision, or positive predictive value.

Table 1
Number of annotations per category in the reference corpus (Set 1 and Set 3 consensus). The number shown between brackets corresponds to reintroduced surrogates for Set 1 and the original de-identified portions for Set 3.

Category	Set 1			Set 3		
	Total	Unique	Ratio	Total	Unique	Ratio
Address	99	29	0.293	89	21	0.236
Zip code	101	17	0.168	97	15	0.155
Date	462 (99)	346	0.749	563 (96)	439	0.780
E-mail	47	19	0.404	64	25	0.391
Hospital	224	61	0.272	224	60	0.268
Identifier	59	48	0.813	76	55	0.724
Last name	871 (148)	409	0.470	985 (146)	342	0.347
First name	750 (131)	255	0.340	834 (117)	190	0.228
Telephone	383	142	0.371	419	157	0.375
City	218	25	0.115	233	29	0.124

These two metrics were selected in order to provide a higher bound (*F*-measure) and a lower bound (κ) for estimating inter-annotator agreement. Grouin et al. [40] have shown that considering all the entities annotated at least once as "markables" provides a lower bound for estimating inter-annotator agreement. Applied to Set 1 in the present study, we thus considered 3213 markables in the overall corpus of 100 documents (i.e., 3213 entities are annotated in the consensus output), 1597 markables in the sub-corpus of 50 documents pre-annotated with the rule-based system and 1616 markables in the sub-corpus of 50 documents pre-annotated with the CRF-based system.

According to Artstein and Poesio [41], agreements higher than 0.8 indicate that annotations can be considered consistent.

3.4.2. Automatic system evaluation

For system evaluation, we used precision, recall and *F*-measure as defined above. The advantage of *F*-measure in this context is that it provides a direct comparison between system performance and inter-annotator agreement.

3.4.3. Categorization and boundary evaluation

The Slot Error Rate (SER) [42] is a composite metric that takes into account both categorization errors (marking the same mention with a different category, compared to the reference) and boundary errors (marking a mention that overlaps with a mention marked with the same category in the reference) (formula (8)). The lower the SER, the better the quality of the corpus.

$$\text{Slot Error Rate} = \frac{D + I + TF + 0.5 \times (T + F)}{R} \tag{8}$$

where "*D*" is the number of deletes (i.e., false negatives), "*I*" is the number of inserts (i.e., false positives), "*T*" is the number of categorization errors only, "*F*" is the number of boundary errors only, "*TF*" is the number of both categorization and boundary errors, and "*R*" is the number of expected elements in the reference (i.e., true positives + false negatives). The numeric value in the formula is a penality which allows weighting of the categorization and boundary errors. In our experiments, we used a penality of 0.5; with this value, the cost of errors increases proportionally with the number of errors.

Fig. 5 provides an annotated example to illustrate the Slot Error Rate metric. The correct annotations are shown in green boxes while errors appear in red boxes. Specifically, only one annotation was correct ("1 avenue de Paris" annotated as an address), one insertion ("Paris" as a city), one deletion ("Paul"), one type error ("Martin" annotated as a firstname instead of a lastname, with correct boundaries) and one boundary error ("Saint Germain" instead of "Saint Germain en Laye" with a correct category). No annotation

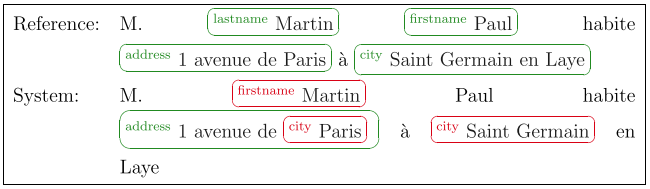

Fig. 5. Fabricated example to illustrate the Slot Error Rate metric.

combines type and boundary errors. The resulting Slot Error Rate is: $\frac{1+1+0+0.5\times(1+1)}{4} = 0.75$.

4. Results

4.1. Evaluation of the reference de-identification corpus

4.1.1. Descriptive statistics

We produced consensus annotations on two 100-document corpora (Set 1 and Set 3). Overall, the distribution of tokens and annotations was comparable in the two corpora. Set 1 contains a total of 29,437 tokens including 3213 tokens (11%) annotated as PHI tokens. The corpus averages 32.13 PHI tokens per document (standard deviation 15.6). The minimum number of PHI tokens per documents is 6 and the maximum is 71 (median 32).

4.1.2. Automatic pre-annotation

The automatic pre-annotation produced by the MEDINA suite of tools was evaluated using the consensus annotations as a reference (Set 1). This evaluation was performed on 50 documents per pre-annotation method because of the annotation methodology (i.e., 100 documents were annotated with 50 documents pre-annotated using a rule-based method and the remaining 50 documents pre-annotated using a CRF approach. Table 2 presents the overall results, depending on the pre-annotation method used to build the consensus. The "In" columns present results of the automatic annotations that were effectively revised by the annotators. The "Out" columns present the results of the automatic annotations obtained with the method that was not revised by the annotators. Table 3 shows the detailed results per category.

The systems used to de-identify the corpus rely on different methods (rule-based vs. machine-learning) and produce distinct annotations. Table 4 shows a classification of the errors produced by each system on the sub-corpus of 50 documents that were then revised by human annotators.

Fig. 6 illustrates the types of errors resulting from the automatic pre-annotation: (i) missing annotation, (ii) category error (here a last name and a first name instead of an hospital name), (iii) spurious annotation (the city name is a part of the street name and should not have been annotated), and (iv) over-anonymization (a word has been annotated while it is not a personal information and the annotation of the correct last name is missing). Both the rule based and CRF-based systems were prone to these errors.

These results obtained on Set 1 guided our decision to use the rule-based system to produce pre-annotations for Set 2.

Table 2
Inter-annotator agreement between the automatic pre-annotation and the human annotations consensus; in = the pre-annotation has been used to build the consensus; out = the pre-annotation has not been used to build the consensus.

	In (50 docs)		Out (50 docs)		All Set 1 (100 docs)	
	κ	F	κ	F	κ	F
Rule-based	0.680	0.813	0.649	0.787	0.662	0.799
CRF-based	0.313	0.519	0.310	0.514	0.311	0.517

Table 3
Detailed agreement between the automatic pre-annotation and the human annotations consensus; the number of entities per category for each sub-corpus in Set 1 is given.

Category	Rule-based (50 docs)		CRF-based (50 docs)	
	#	F	#	F
Address	47	0.667	52	0.038
Zip code	49	0.970	52	0.882
Date	228	0.980	234	0.701
E-mail	29	0.929	26	0.000
Hospital	98	0.201	126	0.210
Identifier	36	0.000	23	0.000
Last name	450	0.746	420	0.555
First name	369	0.852	373	0.679
Telephone	186	0.986	197	0.040
City	105	0.859	113	0.080

4.2. Evaluation of a statistical de-identification tool

Fig. 7 shows the evolution of recall, precision and F-measure on Set 1 depending on the number of documents used in Set 2 for training the CRF model.

These results guided our decision to use the CRF system trained on Set 2 to produce pre-annotations for Set 3.

Table 5 shows the results per category achieved by the CRF model built on the 100-document training corpus (Set 2).

4.2.1. Annotation duration

We computed how much time each human annotator took to revise the pre-annotated corpora of 100 documents in Set 1 and

Table 4
Slot Error Rate for each pre-annotation system used on Set 1; the number of elements within each kind of error is given between parenthesis.

Slot Error Rate	Rule-based 0.255	CRF-based 0.578
Corrects	79.3% (1266)	47.8% (773)
Inserts	8.4% (134)	13.5% (218)
Deletes	13.4% (214)	29.1% (471)
Substitutions	7.3% (117)	23.0% (372)

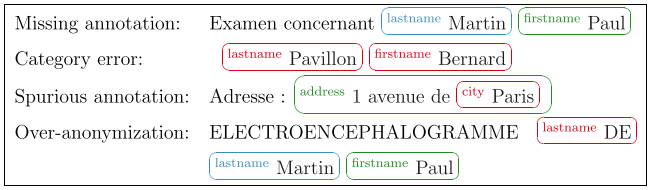

Fig. 6. Sample automatic pre-annotations: correct annotations are shown in a green box, incorrect annotations are in a red box, missing annotations are in a blue box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

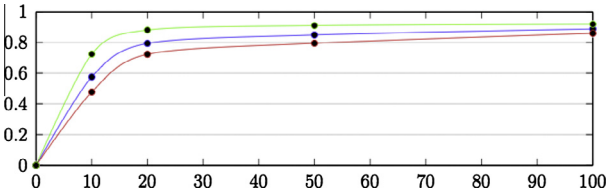

Fig. 7. Evolution of recall (red), precision (green) and F-measure (blue) on Set 1 depending on the number of documents in Set 2 (10, 20, 50 and 100) used to train the CRF models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Set 3. Table 6 shows the time spent by each annotator to revise the pre-annotated documents, overall and by pre-annotation type. An additional 30 minutes period was needed to go over the annotations to arrive at the consensus for Set 1, and 15 minutes for Set 3.

4.2.2. Human annotator agreement

We computed the intra-annotator agreement (Table 7) on the 10 documents in Set 1 that were annotated in duplicate by each human annotator, i.e., based on a version with and without pre-annotations. Table 7 also shows the inter-annotator agreement between both annotators and between each annotator and the consensus annotations on Set 1; recall that the consensus annotations were obtained after the revision of automatic pre-annotations.

The Set 1 reference corpus comprises 100 documents that have been automatically pre-annotated by two systems: a rule-based system for a half (50 documents), and a CRF system trained on outside data for the other half (50 documents). Set 3 comprises 100 documents that were pre-annotated by a CRF system trained on data from Set 2.

Table 8 shows the inter-annotator agreement between both annotators as well as between each annotator and the consensus annotations. The scores were computed on both reference sets for each type of pre-annotation.

Table 9 presents the detailed inter-annotator agreement achieved per category on the overall pre-annotated corpus (i.e., on the 100 documents) and on each sub-part of the corpus depending on the tool used for pre-annotation, either the rule-based or the CRF system. The number of entities per category on the overal corpus is given between parenthesis.

5. Discussion

5.1. Quality of the reference de-identification corpus

5.1.1. Quality of the pre-annotations

Tables 2 and 4 show that the rule-based system performed better than the out-of-the-box CRF system on the reference corpus. The rule-based system achieved higher agreement with the consensus annotations (κ = 0.680 and F-measure = 0.813) and lower Slot Error Rate (0.255) compared to the CRF system (respectively: 0.313, 0.519, and 0.578). More specifically, Table 3 shows that the rule-based system achieved better performance for each category, except for hospital names where the CRF system slightly outperformed the rule-based system. The poorer results obtained with the CRF approach can be explained by the fact that the model has been trained on a corpus of clinical documents that came from a different hospital and belonged to a different medical subdomain than that of test set. The characteristics of both corpora are different, so that the characteristics that have been learnt by the CRF model cannot be found in the test corpus we aimed to de-identify in this study. The CRF model is not robust enough to process correctly a new unknown corpus. However, when training data from the same corpus is available, the statistical model can outperform the rule-based model. Fig. 7 shows a steep progress curve in the performance of the tool as more documents are used for training. The progression slows down after 20 documents are added. The difference in F-measure is 54% between the smaller and the larger training set (from F-measure 0.575 to 0.888). In comparison, the improvement in performance observed by Hanauer et al. [19] between similar size training sets was only 6% (from F-measure

Table 7
(A) Inter-annotator agreement between both annotators (1–2) and between each annotator and the consensus (1-Con; 2-Con). (B) Intra-annotator agreement with and without pre-annotations (1-1raw; 2-2raw), computed using the κ coefficient and the F-measure, on the subset of 10 documents in Set 1 annotated in duplicate.

	Set 1 (100 docs)	
	κ	F
(A)		
1–2	0.917	0.954
1-Con	0.944	0.964
2-Con	0.949	0.974
	Set 1 subset (10 docs)	
	κ	F
(B)		
1raw-1pre	0.877	0.911
2raw-2pre	0.937	0.969

Table 5
Evaluation of the CRF model built on the 100-document training corpus (Set 2).

Category	Set 1 (100 documents)			Set 3 (100 documents)		
	R	P	F	R	P	F
Address	0.818	0.920	0.966	0.888	0.898	0.893
Zip code	0.921	1.000	0.959	0.918	1.000	0.957
Date	0.896	0.935	0.915	0.909	0.979	0.943
E-mail	1.000	0.982	0.991	1.000	1.000	1.000
Hospital	0.737	0.620	0.673	0.786	0.842	0.813
Identifier	0.559	0.943	0.702	0.645	0.961	0.772
Last name	0.797	0.928	0.857	0.954	0.980	0.967
First name	0.880	0.942	0.910	0.966	0.984	0.975
Telephone	0.982	0.977	0.979	0.993	0.990	0.992
City	0.922	0.990	0.955	0.922	0.986	0.953
Overall (micro-average)	0.860	0.919	0.888	0.933	0.973	0.953

Table 6
Revision time in minutes for Sets 1 and 3 (overall time and average time per file), by each human annotator (1 and 2).

	Set 1								Set 3	
	Raw subset (10 docs)		Rule-based (50 docs)		CRF-out (50 docs)		All (100 docs)		CRF-in (100 docs)	
	All	Average	All	Average	All	Average	All	Average	All	Average
1	35	3.50	160	3.20	144	2.88	304	3.04	125	1.25
2	25	2.50	102	2.04	95	1.90	197	1.97	58	0.58

Table 8
Inter-annotator agreement between both annotators and between each annotator (1–2) and the consensus (1-Con; 2-Con), computed using the κ coefficient and the *F*-measure, on Set 1 and Set 3, for each type of pre-annotation used.

	Set 1						Set 3	
	All (100 docs)		Rule-based (50 docs)		CRF-out (50 docs)		CRF-in (100 docs)	
	κ	*F*	κ	*F*	κ	*F*	κ	*F*
1–2	0.897	0.931	0.945	0.964	0.852	0.897	0.959	0.977
1-Con	0.918	0.942	0.968	0.970	0.879	0.914	0.977	0.986
2-Con	0.973	0.984	0.978	0.987	0.967	0.981	0.975	0.987

Table 9
Human inter-annotator agreement in *F*-measure.

Category	Set 1									Set 3		
	All (100 docs)			Rule-based (50 docs)			CRF-out (50 docs)			CRF-in (100 docs)		
	1–2	1-Con	2-Con	1–2	1-Con	2-Con	1–2	1-Con	2-Con	1–2	1-Con	2-Con
Address (99)	0.752	0.767	0.985	0.786	0.797	0.989	0.722	0.741	0.980	0.932	0.932	1.000
Zip code (101)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Date (462)	0.960	0.970	0.989	0.993	0.993	1.000	0.928	0.949	0.979	0.953	0.974	0.979
E-mail (55)	0.953	0.953	1.000	0.966	0.966	1.000	0.939	0.939	1.000	1.000	1.000	1.000
Hospital (224)	0.805	0.806	0.947	0.841	0.830	0.912	0.779	0.787	0.972	0.898	0.936	0.932
Identifier (59)	0.782	0.933	0.840	0.917	0.959	0.958	0.623	0.894	0.700	0.944	0.973	0.945
Last name (870)	0.941	0.956	0.982	0.965	0.980	0.982	0.915	0.930	0.982	0.988	0.993	0.992
First name (742)	0.957	0.965	0.993	0.978	0.984	0.995	0.936	0.946	0.991	0.990	0.994	0.995
Telephone (383)	0.937	0.937	1.000	0.995	0.995	1.000	0.879	0.879	1.000	0.996	1.000	0.996
City (218)	0.941	0.951	0.991	0.986	0.990	0.995	0.897	0.912	0.987	0.991	0.996	0.996

0.888 to 0.942). One possible explanation is that the diversity in the documents in our corpus was higher, therefore more documents (at least 20) are needed to obtain a representative training set. Nonetheless, we observe that using data incrementally has a positive impact as shown by Tu et al. [13] and Hanauer et al. [19].

Using only ten documents for training yields an *F*-measure of 0.575. While low, this performance is still higher than that of the CRF pre-annotation system trained on 230 documents from another healthcare provider, focused on only one medical specialty (*F*-measure 0.517).

The best CRF model, used to pre-annotated Set 3, is obtained with training on 100 documents from the training set (Set 2; see Table 5). It achieves 0.89 *F*-measure on Set 1 and 0.95 *F*-measure on Set 3, which is higher than the state of the art results obtained for Swedish (0.79 [7]) but lower than those obtained for English (0.96 [19]). This can indicate that language may play a role in task complexity. However, direct comparison is difficult because we use different PHIs compared to other work, including the i2b2 challenge [8]. We differentiate between first name vs. last name instead of clinician vs. other, we de-identify dates but not ages, we group different information under the label "identifiers", including patient record number, hospital identification number and medical device serial number. Nonetheless, our results are encouraging in light of the recent finding that de-identification tool performance in the low 90s rivals human performance [14].

When training the CRF system on the training set with and without surrogate PHI re-introduction, we found no difference in performance contrary to Yeniterzi et al. [20].[12] This indicates that our CRF model relies heavily on context information (e.g., neighbor tokens) rather than token characteristics (e.g., is the token in capital letters) to identify PHI.

As in previous research [26,30], we observed that human annotators performed better and faster when pre-annotations were supplied. Annotation time was about 20% faster when revising a pre-annotated corpus compared to annotating the corpus without pre-annotations.

[12] Data not shown.

5.1.2. Annotation time

We expected that annotation time would be consistent with the quality of the annotations: the better the annotations, the less time we expected an annotator to spend on the task. However, Table 6 shows that, for Set 1, both annotators spent about 10% less time revising documents pre-annotated with the CRF system (160 minutes vs. 144 minutes for the first annotator and 102 minutes vs. 95 for the second) even though it performed noticeably worse (*F*-measure 0.519 vs. 0.813). It is possible that the errors in the CRF pre-annotations were more obvious and frequently repeated (for instance, telephone numbers were always annotated as dates) so that annotators were able to quickly revise them, while more careful reviewing of the rule-based pre-annotations was needed to identify errors. Another possible explanation is that some of the CRF errors were overlooked by the annotators and left unrevised, therefore saving annotation time (for instance, punctuation could be erroneously included in the CRF pre-annotations and one of the annotators sometimes failed to correct the annotation span). This explanation is consistent with the lower inter-annotator agreement observed between annotators and between each annotator and the consensus on the documents that were pre-annotated with the CRF system (see Table 8). On Set 3, annotators spent considerably less time revising pre-annotations. These differences can be explained by two factors: first, pre-annotation quality with the custom trained CRF was higher, and second, the annotators acquired training experience working on the previous sets. We also observed the same trend as on Set 1: annotator 2 is faster than annotator 1 regardless of the pre-annotation method.

5.1.3. Quality of the human annotations

Tables 7 and 8 show that the overall inter-annotator agreement is well over 0.80 for both sets, meaning that the annotations are highly consistent. The agreement between annotators and the consensus is also high, which shows that the quality of the annotations is high. These observations remain true when considering the detailed agreement scores per category shown in Table 9. It can be noted that agreement scores are slightly lower for the "hospital" and "address" categories for Set 1, reflecting a difference in

interpreting the guidelines for buildings located inside hospitals. One annotator annotated them as "hospital" and the other as "address" (see Table 8, #3). The consensus enforced the decision to annotate them as "hospital". Therefore, the agreement with the consensus for the annotator who initially chose to annotate them as "address" is lower for these categories. Similarly, consistency is lower for identifiers on the CRF pre-annotated set. This can be explained by one particular case where the guidelines were interpreted differently by the annotators. One annotator annotated ICD and ADICAP codes in the documents as identifiers whereas the other did not (see Table 8, #6); during the consensus process, it was agreed that codes were medical information and should not be annotated as PHI. Only two documents containing a total of 10 ICD codes were impacted, but because they belong to the sub-corpus pre-annotated with the CRF approach, it was enough to create an imbalance in consistency scores for the category. Agreements are generally higher on Set 3 showing that the discussion to create the consensus for Set 1 had a positive impact on the annotators' work for Set 3.

Overall, the scores are quite homogenous over the categories, which means that there is no outstanding annotation difficulty: human annotators were able to consistently identify elements from all the categories. This also shows that the guidelines were defined with an adequate level of detail that allowed annotators to achieve the task with high consistency. Results show that the inter annotator agreement is higher when the quality of pre-annotations is higher. This observation is in line with the results of a curation study showing that gold standard annotations were more helpful than the annotations produced by an NLP tool [43].

Fig. 8 illustrates the most frequent types of disagreements between the human annotators: missed annotations (1; 7), boundary disagreement (2) – and possibly (5), category disagreement (3), over-interpretation of the guidelines (4; 6). Examples of additional difficulties are also shown: distinguishing rare lastnames and firstnames (5) defining identifying codes and numbers (6), and the use of specific abbreviations in the text (7 – here, the name of the hospital). Missed annotations and boundary or category disagreements account for the bulk of annotator disagreements.

Similarly to Rosset et al. [30], we also noticed that the feeling of the human annotators sometimes differed from factual observation. For instance, the annotators correctly realized that one of the pre-annotation methods performed better than the other. However, they thought that they spent less time revising documents pre-annotated with this method which was not true.

#	Annotator 1	Annotator 2
1	Pr Paul [lastname Martin]	Pr [firstname Paul] [lastname Martin]
2	il serait bien que M. [lastname D.] voie rapidement un cardiologue	il serait bien que M. [lastname D.] voie rapidement un cardiologue
3	Examen pratiqué au [address Pavillon Bernard]	Examen pratiqué au [hospital Pavillon Bernard]
4	le [date jeudi] [date 7 mai 2013]	le jeudi [date 7 mai 2013]
5	Examen concernant [firstname Belart] [lastname Angilbe]	Examen concernant [lastname Belart] [firstname Angilbe]
6	ADICAP : BHFF0110	ADICAP : [identifier BHFF0110]
7	Sortie de pneumologie [hospital SN] le [date 10/11/12]	Sortie de pneumologie SN le [date 10/11/12]

Fig. 8. Comparison of annotations performed by the human annotators: correct annotations are in a green box, incorrect ones are in a red box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.1.4. Sources of annotation inconsistencies between human annotators

Inconsistencies were due to (i) distraction (e.g., lack of modification or removal of an erroneous pre-annotation, selection of an erroneous category for an annotation), (ii) pre-annotations (occurrence of full stops in initials, occurrence of vertical bars in telephone numbers), and (iii) variation in interpretation of the guidelines (e.g., annotation of buildings within the hospital as hospital vs. address; annotation of ICD codes as identifiers). For the first two types of inconsistencies, no discussion was needed during the consensus process; usually both annotators recognized that a mistake was made by one of the annotators. For the last type, a discussion was needed to decide how to interpret the guidelines and apply them consistently throughout the consensus process.

5.2. Recommendations for developing de-identified corpus

Based on this study, our recommendation for future corpus development using the MEDINA suite of tools would be to work incrementally according to the following steps: (1) Pre-process a small corpus (e.g., 50 documents) with the rule based system, (2) review, (3) train the CRF system on the available corpus, (4) use to pre-process additional corpus, (5) review, repeat steps 3–5 until desired corpus size is reached. The review steps should systematically involve at least two annotators until inter-annotator agreement reaches a high plateau, e.g., above 95% *F*-measure. The use of re-introduced surrogates in training data has no impact on the performance of our statistical system (data not shown). It produces authentic-looking text, which makes false negatives less noticeable [44]. For this reason, it may be best to introduce the surrogates after a de-identification consensus has been reached, and before the corpus is shared or used for other research purposes.

5.3. Limitations of this study

There are a few limitations to this study, that reflect the complex task of producing an annotated reference corpus. First, the definition of some categories may be too broad, making them difficult to annotate. For instance, the identifiers category covers both medical record numbers, healthcare providers identification numbers and hospital ward numbers. Similar issues were observed during the creation of the Swedish reference corpus, and some category definitions were changed to improve the quality of the annotations [7].

Furthermore, the authors defined the annotation guidelines, worked on the human revisions and the final consensus. While both have extensive annotation experience, it is always desirable to have an outside independent party validate the annotation decisions made. Osman et al. [45] show that for opinion assessments, inter-annotator agreement can vary by a wide margin depending on the annotator pairs. We believe that PHI identification is a task that is inherently less subjective than opinion assessment, so that inter-annotator agreement should be more robust over different pairs of trained annotators. Nonetheless, this advocates the use of additional annotators to validate the gold standard.

It can also be argued that annotating documents without pre-annotations before documents with pre-annotations makes it difficult to specifically attribute annotation time gains to practice or the pre-annotations. However, previous work that focused on showing that pre-annotations save time designed experiments that avoided this caveat [30]. Our main focus here was to compare the use of two types of pre-annotations.

5.4. Future work

In future work, we would like to explore further the evolution of the statistical system performance as more training data is used. It would also be interesting to study methods of selecting the documents included in the training data using active learning, instead of random selection. System performance may also be improved using a hybrid method that would take advantage of the rule-based performance on some categories such as zipcodes, emails and dates. We also plan to apply the protocol defined in this study to the systematic development of a large de-identification reference corpus for French. The corpus will eventually be released to the scientific community.

6. Conclusions

In this paper, we presented several experiments on the de-identification of French clinical notes in order to design a protocol to build a reference corpus optimizing time, annotation quality and annotator experience. We worked sequentially with three sets of 100 documents randomly selected from a group of French hospital. The documents were partially de-identified for the most sensitive data (i.e., first name, last name and date of birth of patients).

First, we used two distinct de-identification systems designed to process clinical notes in French from a cardiology ward: a rule-based system and a statistical system (that relies on the CRF formalism). Then, we compared two CRF models: an out-of-domain model, that has been built on data from another hospital in another medical domain, vs. an in-domain model, built on the result of the human annotation process we designed.

During all experiments, two human annotators revised the result of automatic pre-annotation obtained either from the rule-based system or the machine-learning approach. We computed annotation time as well as inter- and intra-annotator agreement to assess the support provided by the pre-annotations.

In this study, we found that the rule-based system provided better annotation quality compared to the CRF model, when both systems are trained on outside data. We also found that a CRF model built on in-domain data outperformed all other methods. We showed that human annotators worked faster with the pre-annotations obtained from a CRF approach, compared to a rule-based system, even if the CRF-based annotation quality was worse.

This study contributes to research in de-identification by providing unique data on French. It also contributes to research in corpus annotation and reference corpus development by providing insight on how to use available pre-annotation methods to optimize annotation time, annotation quality and annotator experience.

Acknowledgments

This work has been supported by grant CABeRneT ANR-13-JS02–0009-01. We used the scoring tools developed by Dr. Olivier Galibert (LNE, Trappes, France) as part of the Quaero project. The authors thank Dr. Pierre Zweigenbaum for his insightful comments on this work.

References

[1] Meystre S, Savova G, Kipper-Schuler K, Hurdle J. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128–44.
[2] Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care 2012;50(7):S82–S101.
[3] Meystre S, Friedlin F, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010;10:70.
[4] Ruch P, Baud R, Rassinoux A, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. In: Proc AMIA symp; 2000. p. 729–33.
[5] Grouin C, Zweigenbaum P. Automatic de-identification of French clinical records: comparison of rule-based and machine-learning approches. Stud Health Technol Inform 2013;192:476–80.
[6] Velupillai S, Dalianis H, Hassel M, Nilsson G. Developing a standard for de-identifying electronic patient records written in swedish: precision, recall and *f*-measure in a manual and computerized annotation trial. Int J Med Inform 2009;78:e19–26.
[7] Dalianis H, Velupillai S. De-identifying swedish clinical text – refinement of a gold standard and experiments with conditional random fields. J Biomed Semantics 2010;1:6.
[8] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14(5):550–63.
[9] Névéol A, Grouin C, Darmoni SJ, Zweigenbaum P. Désidentification d'un corpus clinique pour le traitement automatique du français. In: Actes de la Session francophone à MedInfo, Copenhagen, Denmark; 2013.
[10] Neamatullah I, Douglass M, Lehman L, Reisner A, Villarroel M, Long W, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.
[11] Saeed M, Lieu C, Raber G, Mark R. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol 2002;29:641–4.
[12] Lee J, Scott D, Villarroel M, Clifford G, Saeed M, Mark R. Open-access MIMIC-II database for intensive care research. In: Proc IEEE eng med biol soc; 2011. p. 8315–8.
[13] Tu K, Klein-Geltink J, Mitiku TF, Mihai C, Martin J. De-identification of primary care electronic medical records free-text data in Ontario, Canada. BMC Med Inform Decis Mak 2010;10:35.
[14] Deléger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013;20(1):84–94.
[15] Ferrández O, South B, Shen S, Friedlin F, Samore M, Meystre S. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Med Res Methodol 2012;12:109.
[16] Ferrández O, South B, Shen S, Friedlin F, Samore M, Meystre S, et al. A best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013;20:77–83.
[17] Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc 2007;14(5):564–73.
[18] Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE identification scrubber toolkit: design, training, and assessment. Int J Med Inform 2010;79(12):849–59.
[19] Hanauer D, Aberdeen J, Bayer S, Wellner B, Clark C, Zheng K, et al. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. Int J Med Inform 2013;82(9):821–31.
[20] Yeniterzi R, Aberdeen J, Bayer S, Wellner B, Hirschman L, Malin B. Effects of personal identifier resynthesis on clinical text de-identification. J Am Med Inform Assoc 2010;17(2):159–68.
[21] Grishman R, Sundheim B. Message understanding conference-6: a brief history. In: Proc of coling, ACL, Stroudsburg, PA; 1996. p. 466–71.
[22] Bossy R, Jourde J, Manine A-P, Veber P, Alphonse E, van de Guchte M, et al. BioNLP shared task – the bacteria track. BMC Bioinformatics 2012;13(Suppl. 11):S3.
[23] Lu Z, Kao H, Wei C, Huang M, Liu J, Kuo C, et al. The gene normalization task in BioCreative III. BMC Bioinformatics 2011;12(Suppl. 8):S2.
[24] Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514–8.
[25] Doğan RI, Lu Z. An improved corpus of disease mentions in PubMed citations. In: Proc of BioNLP; 2012. p. 91–9.
[26] Névéol A, Doğan RI, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. J Biomed Inform 2011;44(2):310–8.
[27] South BR, Shen S, Barrus R, DuVall SL, Uzuner O, Weir C. Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. In: Proc AMIA symp, Washington, DC; 2011. p. 1243–51.
[28] Thiessard F, Mougin F, Diallo G, Jouhet V, Cossin S, Garcelon N, et al. RAVEL: retrieval and visualization in electronic health records. In: Stud Health Technol Inform 2012;180:194–8.
[29] Health insurance portability and accountability act, §164.514; 1996. <http://www.hhs.gov/ocr/AdminSimpRegText.pdf>.
[30] Rosset S, Grouin C, Lavergne T, Ben Jannet M, Leixa J, Galibert O, et al. Automatic named entity pre-annotation for out-of-domain human annotation. In: Proc of LAW-VII & ID, ACL, Sofia, Bulgaria; 2013. p. 168–77.
[31] Grouin C. Anonymisation de documents cliniques: performances et limites des méthodes symboliques et par apprentissage statistique, PhD thesis. University Pierre et Marie Curie, Paris, France; 2013.
[32] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proc of ICML, Williamstown, MA; 2001. p. 282–9.
[33] Lavergne T, Cappé O, Yvon F. Practical very large scale CRFs. In: Proc of ACL, Uppsala, Sweden; 2010. p. 504–13.
[34] Brown PF, Della Pietra VJ, de Souza PV, Lai JC, Mercer RL. Class-based n-gram models of natural language. Comput Linguist 1992;18(4):467–79.

[35] Liang P. Semi-supervised learning for natural language. Master's thesis. Massachusetts Institute of Technology; 2005.
[36] Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. In: Proc of EACL demonstrations, ACL, Avignon, France; 2012. p. 102–7.
[37] Neves M, Leser U. A survey on annotation tools for the biomedical literature. Brief Bioinform 2012.
[38] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37–46.
[39] Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge, MA: MIT Press; 2000.
[40] Grouin C, Rosset S, Zweigenbaum P, Fort K, Galibert O, Quintard L. Proposal for an extension of traditional named entities: from guidelines to evaluation, an overview. In: Proc of LAW-V, ACL, Portland, OR; 2011. p. 92–100.
[41] Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist 2008;34(4):555–96.
[42] Makhoul J, Kubala F, Schwartz R, Weischedel R. Performance measures for information extraction. In: Proc of DARPA broadcast news workshop; 1999. p. 249–52.
[43] Alex B, Grover C, Haddow B, Kabadjov M, Klein E, Matthews M, et al. Assisted curation: does text mining really help? Pac Symp Biocomput 2008:556–67.
[44] Carrell D, Malin B, Aberdeen J, Bayer S, Clark C, Wellner B, et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. J Am Med Inform Assoc 2013;20(2):342–8.
[45] Osman D, Yearwood J, Vamplew P. Automated opinion detection: implications of the level of agreement between human raters. Inf Process Manage 2010;46(3):331–42. http://dx.doi.org/10.1016/j.ipm.2009.08.005.